A Study of the Effect of Context and Test Method in Evaluating Safety Symbols

by Jennifer Snow Wolff

In Response to the five yearly revision of the
ANSI (American National Standard) z535.3-1991
Criteria for Safety Symbols

Submission of Revisions of the Annex A (Normative)
Suggested Procedure for Evaluating Candidate Symbols

TABLE OF CONTENTS

Appreciations

My greatest thanks go to Dr. Michael Wogalter, for introducing me to experimental design methods in 1990, when I was simply a curious designer working in Albany, NY, and he was a professor of human factors at RPI in Troy, NY. He helped me through our first collaborative research and paper. He encouraged me to apply to graduate school and wrote recommendations. He found funding for this study, and believed enough in me to give it to me, although I was in Atlanta, Georgia and he was teaching at NCSU in Raleigh, North Carolina. His contribution as a member of my thesis committee was extensive and detailed in every phase of design and writing, despite being mediated by telephone, fax and computer e-mail. Dr. Wogalter has been encouraging, patient and cheerful throughout our interactions. I look forward to further collaborative work.

Thank-you to the other members of my thesis committee at Georgia Tech. Dr. William Evans, not only my thesis advisor, has been supportive during my entire time at Georgia Tech. Dr. Neff Walker provided some excellent suggestions about the experimental design. Dr. Joseph Petraglia provided much needed criticism. Thanks also to Dr. Karen Jost of the University of Georgia.*

Thank-you to Blair Brewster of ElectroMark Corporation for his input, technical resources, and financial support of this research.

Thank-you to my advisor, Dr. Peter McGuire, for believing in me, encouraging me and funding me for two years at Georgia Tech.

Thank-you to my mother, Dr. Bettina Wolff, for her suggestions, encouragement, editing skills and much more.

Thank-you to the Graphics, Visualization and Usability lab, Jim Foley, Randy Carpenter and Elaine Swobe for their help making this project into a technical report, and for the use of the GVU facilities.

1. A Study of the Effect of Context and Test Method in Evaluating Safety Symbols

I. ABSTRACT

The study measured the effect of context and test method in evaluating safety symbols. The study consisted of a 2x3 factorial test with context and no context as one independent variable and plausible and poor multiple choice distractors and open ended testing methods as another independent variable. Thirty-three symbols were tested across all six conditions.

The study measured the effect of the quality of multiple choice distractors or alternative answers on scores. The open ended comprehension method was used as a control to measure the ability of multiple choice to capture participant responses. It was found that typical distractors obtained in an independent, seemingly valid test were below average in plausibility compared to distractors obtained through open-ended comprehension testing. Furthermore, it was shown that the low plausibility of those distractors, and its corresponding limit in range of allowable answers led to inflated percentage of correct scores. The average difference between low and high plausibility distractors was 30% across all 33 symbols.

Providing pictorial context in the test environment resulted in a more valid method of raising symbol scores. Context, in this study, was provided by 1 to 4 color photographs of probable environments where a symbol would appear. Contextual cues in the symbols were defined as environmental detail such as water, moon, stars, building structures and identifiable tools or machines. Context effects in testing were found for simpler symbols low in contextual cues in the symbols themselves. Context effects were not found in symbols which contained contextual cues or detail. The average effect of context manipulation for (the simpler, low context symbols) in the open-ended comprehension testing method was an increase of 15 percentage points. The average effect in the multiple choice method was an increase of 18 percentage points in the good distractor condition and 7 in the low plausibility distractor condition.

The ability of context to raise scores is important because a valid method of testing which will also result in symbols which can exceed the ANSI 85% standard. This is important to producers of hazardous products for liability protection and because performing better will reduce the costs of developing pictorials. Providing empirical proof of symbol comprehension is a critical feature for safety and potential litigation. The principle issue in products liability cases involving warning defects is whether the product failed to contain an adequate warning about the dangers inherent in using the product. (Grisim, 1993)

The findings of the study are also important, because it shows outcome of testing varies as a function of test method and the materials (context) provided during the test. They study also suggests that some of the scores which placed symbols currently in the standard were invalid. Furthermore, it suggests that the inclusion of the multiple choice method, a commonly used method of symbol testing in laboratories across the country should be removed from the standard.

II. INTRODUCTION

The study was driven by practical considerations and real world constraints: those of providing the American National Standard with a usable testing method that is both

valid and will provide symbols that are comprehensible to most, if not all, of the intended target population.

The study was designed to measure the validity of the multiple choice method of testing pictorials, which is currently a recommended testing method in the standard. The hypothesis is that low plausibility distractors will result in artificially high percentage of correct scores for pictorials on the multiple choice tests. Additionally, the focus of the study was to measure how difficult or unlikely it is to obtain plausible distractors and to measure the extent to which multiple choice can fail to obtain valid results.

The open ended comprehension method was run concurrently under the same conditions with the same test materials to measure the validity of the multiple choice method results.

To provide measurable evidence, and in order to forestall objections that less plausible distractors would never occur in a real testing condition, two methods of selecting distractors were used. The majority (73%) of the symbols tested and their distractors were taken from an earlier pictorial study that used a multiple choice test (Collins, 1983). Additional distractors were obtained from test participants in an open-ended written comprehension test. The plausibility of all distractors were measured by test participants on a 7 point Likert rating scale.

Context was included in the form of 1 to 4 photographs showing where a symbol might appear to measure its effect on comprehension scores. There was an average 10% increase in correct comprehension scores across the board, with a 16.8% difference for symbols which contained little contextual (or environmental) detail imbedded in the symbol itself. Symbols which contained contextual detail showed little or no effect of context manipulation.

The plausibility scores for the Mining study distractors were below the average for all of the distractors tested. Thus, there is the possibility that without elaborate measures to obtain better ones, (such as conducting a preliminary open-ended test as was done in this research) the quality of distractors in a typical multiple choice test could be very poor.

III. METHODS

1. Part One: Initial Selection of Symbols and Distractors

Initially, a set of 39 pictorials were chosen. Six were ultimately discarded because of similarity to other symbols, or because six distractors couldn't be found, so that the final set tested was 33.

A 4 multiple choice test was used which contained one correct choice and three distractor choices. Thus, 6 valid multiple choice distractors, (three high and three low plausibility) would be needed. Therefore, symbols were chosen which had been tested previously so that independently derived distractors would be available. Twenty-eight of the symbols came with both previous independently derived scores and a set of verifiable wrong interpretations derived from actual test participants.

Nine pictorials came from tests of pharmaceutical pictorials conducted by Wolff and Wogalter that had been tested in the open ended method. These symbols had many more than three wrong interpretations to use as distractors which had been identified by the previous open-ended research. Seven of those were finally selected.

Twenty-five pictorials came from a Mining Hazards multiple choice study published in a technical report in 1983 by Belinda Collins under the auspices of the U.S. Department of the Interior. Some of these symbols are shown as part of the 1991 z535.3 standard. Each of these symbols had 3 distractors that had been used in that study.

These distractors had to be tested for plausibility and an additional three distractors for each needed to be derived for our study so that three plausible and three implausible were assembled. Twenty-one of these symbols made it to the final study.

A third group of five symbols were obtained from a variety of published articles so that the study would be generalizable to many classes/categories of pictorials. Several of these symbols had been tested previously and identified as well or poorly understood. Several other symbols had undergone no testing. Six distractors had to be derived for these 5 symbols.

In the few cases where independently derived distractors were not available, additional initial distractors were derived informally so that each symbol had six distractors for the initial plausibility rating.

2. Part Two: Derive Plausibility Ratings and Additional Distractors

The second phase was to derive a plausibility rating for each of the 6 distractors of the 33 possible test pictorials. Three plausible and three low plausibility distractors would be needed for the final test.

A preliminary set of potential distractors needed to be assembled for testing. Using previous research and common sense, a set was assembled of 274 distractors that ranged from high to low plausibility.

The first plausibility study was run in the Georgia Tech Student Center with 75 student participants. Each of the candidate distractors and the 33 correct referents was rated for plausibility on a 7 pt. Likert scale. None of the 75 participants saw the same symbol twice. Each distractor was scored 10 times by a different evaluator. Although a total of 2,400 scores were analyzed, only 5 of the 34 symbol candidates yielded 3 distractor scores above the median.

To obtain even more plausible distractors, a separate open-ended comprehension study was run with 100 participants in two separate locations, a contra dance and a circus. Participants wrote their understanding of a subset of half of the 34 symbols. By selecting distractors from these responses, an additional 78 candidate distractors were obtained for further plausibility testing.

A second plausibility study was run with an additional 50 participants in Piedmont Park. Participants were given candy or sodas for their participation.

When the results of the second plausibility study was compiled, additional distractors were chosen for the final test. In the final test, only 6 symbols had 3 distractors with ratings over the median score of 4.0, 12 symbols had only 2 plausible distractors and 8 had only 1.

Standard criteria for choosing distractors in multiple choice tests was researched (such as not choosing overlapping concepts) and followed when assembling distractors.

Materials

Two pages of the text booklet were placed on one 8 1/2 by 11 sheet of paper, resulting in 108 pages or 216 different possible combinations. 10 copies were made of 108 pages. The papers were cut in half and collated into 7 different piles, with a different distractor in each pile. The 7 different distractors were randomly placed in piles to avoid order effects. Each pile was then shuffled and stapled together to reduce further any order effects.

Procedures

The first plausibility test was held in the Student Center of the Georgia Institute of Technology. The second was held in Piedmont Park. The test took 3 to 7 minutes. Each participant was allowed to choose from a variety of candy, crackers and gum on the table.

Each participant received a pen and paper test packet containing each of the 39 pictorials and one of 6 to 10 possible distractors or the correct answer written underneath it. A seven point rating scale appeared below the text and the pictorial.

Participants were given the following instructions.:

"This is a packet of pictures, each picture is different and has a different line of text below the picture. The scale below is the same on each page. (Researcher ruffles through the packet quickly to show them) Please look at the picture and read the text. If you think the picture matches the text perfectly circle the 7, if you think that it is completely wrong, circle the 1. You may pick any number on the scale from 1 to 7 if you think it is somewhere in between. Please look at the scale now, and read it carefully. Once you have read it a few times and know it, the test should go very quickly. If you've seen the picture before or there is a second of the same picture, just cross it off."

Participants

123 participants completed the surveys. They ranged in age from 17 to 56. The average age was 24 with a standard deviation of 6.5. 87 of the participants were male and 36 were female. 112 spoke English as their main language, 2 spoke Chinese, 3 spoke Spanish, 4 spoke other languages such as Russian or Turkish, and 2 were unknown.

3. Part Three: Assemble Context Materials

A set of color photographs that provided external context for each symbol was developed for use in the context condition of the study.

Glossy color photographs were obtained from catalogs and magazines retrieved from a commercial garage (automotive, and boating trade magazines and tools catalogs) and a dumpster (architectural, scientific, sports and news magazines). Additional photographs were obtained by shooting and developing original photographs at sites on the Georgia Tech campus (an eyewash and first aid station, chemistry apparatus, exit doors, machine tools, cylinders, pipes and electrical boxes). A selection of 1 to 4 photographs were selected to represent a cross section of the different environments where a symbol might be placed.

A 2 inch copy of the symbol, its corresponding number to match the answer sheet along with the photographs showing where the symbol might appear were assembled on an 8 1/2 by 11 sheet of paper. Six sets of color xerographic copies of the 33 symbols were slipped into plastic sheet protectors.

These sets were used for the context in both the open ended and multiple choice context conditions.

4. Part Four: The Main Study

This phase of the study consisted of 2x3 factorial study with context and no context on one axis and plausible and poor multiple choice distractors and open ended testing methods on the other.

Average Scores for All Symbols

	Multiple Choice		Open-ended
	Plausible	Bogus	
No Context	**59.8%**	**88.1%**	**55%**
Context	**69.8%**	**92.6%**	**63.6%**

Test booklets were prepared for each method and administered in each condition simultaneously with the same populations. The materials were identical across both context and non-context conditions.

Multiple choice distractor scores obtained in the first two phases of the study were used in the multiple choice test booklets. When available, distractors with scores of at least 4.0 were used in the "good" multiple choice condition, and scores no higher than 3.0 were used in the low plausibility multiple choice condition. If this condition could not be met, the highest or lowest available scores were used in the two conditions.

Materials

Each test booklet contained a covering instruction sheet, a biographical data sheet and a Georgia Tech Human Subject Informed Consent Release form.

Four separate instruction sheets were developed and copied onto different color paper. The instruction sheets for the high and low plausibility versions of each of the multiple choice conditions were identical with the exception of a small identifying number on the test pages.

The open ended answer sheets consisted of 3 pages numbered consecutively from 1 to 33, with eleven sets of three blank lines next to the numbers on each page.

The participants in the no context condition were given a small 3.3" by 8 1/2" booklet with only the symbol and its corresponding number (printed in 48 pt. bold type) on each page. The pages of the booklet were held together with a clip and shuffled randomly after each participant to reduce order effects.

The participants in the context condition were given the set of color Xerox pages which was held together with a clip and randomly shuffled for each participant.

The participants in the four multiple choice conditions all received 5.5' by 8 1/2' answer booklets. Each booklet consisted of 33 pages that had a symbol, its corresponding number and 4 multiple choice answers.

The participants in the no context condition received a booklet that had been shuffled randomly and stapled together. They were told to follow the order of the answer booklet.

The participants in the context condition received an answer booklet arranged in numerical order. They were also given the same set of color images used in the open ended context condition. They were told to follow the random order of the photographs, to find the matching number in their test booklet and choose the answer there.

Procedures

The tests was held in a busy hallway of the Student Center of the Georgia Institute of Technology, at the Atlanta Friends Meeting House and The Ponce de Leon Senior Citizens Center. The tests took from 5 to 25 minutes, with the multiple choice method, no context method taking the least amount of time and the open-ended, context condition taking the most time.

Each participant was given candy bars, money, a baseball cap or a donation to their church or senior center in return for their participation in the study.

Each participant was verbally given instructions in addition to the written instructions. The participants were told that they should write or select the meaning that the symbol communicated. They were not told that the symbols were warning or safety symbols, they were referred to only as symbols. The No Smoking symbol was used as an example on the cover of the test booklet.

Participants in both the open-ended conditions and the context/multiple choice conditions were told to write their answers in the order of the random symbols, not the numerical order of their answer booklet. All participants were told not to go back and change their answer.

Participants

The ISO recommends over-representation of participants under 30 and over 50. Students at Georgia Tech provided the under 30 participants. An additional group of older participants were found at a church and a senior citizens center. In addition, the Georgia Tech student population provided a diverse mixture of races.

Overall, the participants were 57% male, 42% female and 1% unknown. 66% of the participants were under 30 and 18% were over 50. 47% of the participants had some college, 16% were college graduates and 21% had done postgraduate work. 10% of the participants listed a language other than English as their first language. 60% of the participants listed student as their job, 9% as retired, 3% as homemakers, and the others in a variety of occupations.

Responses given by participants in Part 2 (from the circus and contra dance) were not compiled with the main study results because it could not have been determined whether differences were due to population difference, a different time, setting or other factors rather than test methodology. Those results were only used to gather distractors for Part 4. That data was not compiled, but might be at a later time.

Scoring

The open-ended answers in the context and no context condition were scored by two independent judges who were paid $7.00 an hour for their time. The judges did not meet, and were only given the correct meaning of the symbol and the verbal written answers of the test participants. They did not see the symbols before scoring, so that their responses were not biased by the image. This method revealed that the verbal label for some of the symbols were not clearly written (see discussion section on poison hazard).

To prevent fatigue or learning affecting the judge's responses, the participant's responses for the context and no-context versions were collated together. In order to measure the occurrence of the distractors in the open-ended responses, a scoring sheet was created for the use of the judges. The scoring sheet had the correct response and each of the distractors. The correct answer appeared at the top of the form, then the 6

distractors and spaces for a blank answer, an "I don't know" answer and an "other" category. The judge was asked to match the participant response to one of the answers.

Each judge received his or her own scoring sheets and the responses were entered manually into the computer later. The incorrect responses were used only to determine how often the distractors occurred in an open-ended test, and to determine the range of answers that occurred in the open-ended test as compared to the multiple choice test.

However, in determining the judges inter-rater reliability, agreement was considered to be both judges choosing the correct answer or both judges choosing any incorrect answer. A disagreement was considered to be found only if one judge found the answer to be correct and another judge found the answer to be any incorrect response.

One judge was a Master's candidate in the Information, Design and Technology Program at Georgia Tech, a married female in her thirties, a Canadian citizen with data entry experience. The second judge was a male in his fifties, a freelance writer, theology student, with a degree in psychology, born and raised in Georgia.

IV. BACKGROUND RESEARCH AND THEORY SUPPORTING THE RECOMMENDATIONS AND EXPERIMENTAL DESIGN

1. The open-ended method cannot be adequately replaced by multiple choice

The open-ended comprehension method is currently mentioned in the standards, but it is not emphasized that it is a superior method. Robert Dewar states that the open-ended testing procedure they call comprehension testing "shows the extent to which symbols are understood correctly and is therefore the most important testing procedure in the development of public information symbols." He reiterates the importance of "qualitative data" that helps designers improve variants. ISO has such confidence in this method that it is used as a benchmark to measure the validity of the many preselection, ranking and estimation methods they have considered. (Brugger, 1994)

2. Research suggests that the multiple choice method can and should be eliminated

Jennifer Snow Wolff has recently completed independent masters research to provide this committee with empirical evidence as to the validity of the multiple choice method. A detailed description of this research can be found in the methods section (Wolff, 1995).

However, the Wolff research shows that the multiple choice method is highly dependent on the quality of distractors. The process of identifying high quality distractors that could match the validity of the comprehension method was more time and cost intensive than the comprehension method alone.

Furthermore, it was shown by analyzing the results of a published ANSI report which used the multiple choice method that the method failed to identify a number of poor icons. The quality of the distractors and the method for selecting them was the probable cause of that failure.

For example, in the Mining Study (Collins, 1983), distractors in a multiple choice test had been in response to a different variant of the symbol being tested. In relation to the current variant, the distractors were not plausible. An example of this is the slip hazard symbol (9). The distractors, "Keep area clean," "Wear boots in area," and "dangerous poisonous snakes in area" were in response to an earlier version of the

symbol that included a large boot and a wiggly line. Although in this case, the symbol is a good one, had it been a poor symbol, these answers would have caused this symbol to pass unfairly. In order to avoid this problem, each and every symbol variant would have to undergo comprehension testing to identify plausible distractors for that variant. This eliminates any ease of testing feature that made the multiple choice test a method of choice, for some test makers.. It can be seen that there is no way to validly replace comprehension testing with multiple choice.

Furthermore, Robert Dewar (Dewar, 1994), an experienced and well respected symbols researcher has also published concerns about the multiple choice method, and his opinion stated here, concurs with the recommendations of this report.

> "The ease of understanding is perhaps the most important single index of a symbol's effectiveness. One of the best ways to measure this is to show participants a photograph or slide of the sign (preferably in context) and have them write in an answer booklet the symbol's meaning. Data reduction is time-consuming, but the extra effort pays off in terms of a wealth of information about the types of errors and confusions people make, and may assist in the redesign process. While multiple choice methods are more efficient, it is often difficult to select appropriate wrong responses and minimize the effects of guessing."

3. The use of photographic reference can eliminate out of context answers

A number of answers that participants give in laboratory tests result from the the testing environment. When no contexual clues are supplied, the participant will supply their own, (also known as perceptual set) which may or may not be coming from clues in the symbol itself. For instance, one participant at the circus thought a symbol had something to do with the circus. In a real world situation, context is usually supplied by the setting. Photographs or a verbal description can provide more valid contextual clues in a laboratory testing environment.

The already cited symbol expert, Robert Dewar points out that

> "it is too costly to conduct field evaluations of symbols. A more efficient and much less expensive approach is to evaluate them in the laboratory. ...It is essential to ensure that [laboratory testing] methods are properly validated against 'real world' measures of the effectiveness of the symbols in *context.* Unfortunately, there has been a tendency to accept these methods without properly validating them against appropriate criterion measures. ...Most of the research on comprehension of symbols is done in laboratory or classroom settings using visual material that does not convey any information about the context in which the symbol might appear."

The inclusion of pictorial reference materials or verbal scenarios to provide some "context" is just a first step towards the goal of validation against real world measures. A number of researchers have looked at two forms of replicating contextual information in the laboratory, a written verbal message, and photographic images.

Vukelich and Whitaker (1993) examined the importance of context in symbol comprehension by providing participants with a more elaborate context (a two-sentence description of the setting in which a symbol might be seen), partial context (a two-word description of the context), or no context information. An example of a more elaborate context for the 'lost and found' symbol would be 'You are walking in an international airport. This symbol is located on a sign extending from the wall overhead.' Comprehension was higher with the more elaborate verbal description. Familiarity with the symbols also enhanced their comprehension.

Another study by in 1975 by Cahill tested ten graphic symbols designed by the well-known Henry Dreyfuss in context and in isolation. Symbols were more correctly identified in context and by participants with prior relevant experience.

Cahill's context consisted of a drawing of the interior of a cab, and additional verbal instructions, and numbers corresponding to the symbols shown. The symbols were shown for 5 second intervals in a darkened room. The symbols were scored as correct or incorrect, and the experiment used three independent judges who did not know if a test was done in the context or no-context situation. Agreement of at least two judges was required for the item to be judged correct.

Both Brelsford, Scoggins and Wogalter (1994) and Silver, Wogalter, Magurno and Glover (1995) assessed the comprehensibility of warning pictorials given the presence or absence of context. Context was provided by a photograph and a verbal description of an environmental scene. In the latter study, the sample was divided into context and no context conditions on pictorial variants from three categories, Keep Out, Electrical Danger and Do Not Dig. Then, the participants were told what the pictorials meant and rated all five pictorials in each category on quality and then ranked the five pictorials for effectiveness of the message. The context manipulation showed differences in comprehension for only one pictorial category, the Electrical Danger, and showed no difference for the other two. A high positive correlation was obtained for the ranking and rating procedures.

In the Wolff study (1995), differences in effect for context were seen only in simpler pictorials which did not contain a lot of environmental details in the symbol itself (such as a door, a bed, equipment such as trucks, boats, bodies of water, sky etc.)Thus the context was ambiguous. (This result may explain lack of context effects in the Silver study, whose symbols contained external environmental details in the symbols themselves.)

4. The introduction of iterative pre-testing and design

Evaluation of pictorials can be considered to be part of an ongoing, iterative design process, rather than a final testing procedure. There is a good deal of evidence recommending this approach. In the evaluation of graphical computer interfaces, usability testing takes this approach with great success. Iterative testing and redesign results of 4 design teams which were given to a fifth design team resulted in a 200% improvement in usability of the graphical interface (Nielson, 1993). The process uses a small number of carefully chosen, naive, participants (e.g. 5), the collection of quantitative and qualitative data, and many iterations in the design process. Usability testing found that the new information yield was minimal after the first 5 participants. ((Kraemer, H.C. & Thiemann, S. (1987), Lewis, J. R. (1994), Nielson, J. (1993), Virzi, R. A. (1990, 1992) The use of iterative design of symbols with smaller number of participants can yield better final symbols, forewarn researchers of problems understanding the symbol or its referent and thus ease the problems in final testing and evaluation.

I cannot recommend eliminating the use of larger number of participants, based solely on the experience of usability studies. However, it is an effective methodology which can and should be explored for effectiveness in the domain of symbol testing. In usability testing, there is a great deal of information that can be gathered from one participant, and sessions typically last several hours for one participant. The emphasis here is on the iterative design of the interface. Perhaps smaller numbers of participants can be used effectively in the iterative design stage, and larger numbers used in the final comprehensive testing stage.

5. The inclusion of a preliminary ranking or estimation phase

Symbols researcher, Brugger, notes that the rigorous testing procedure defined in ISO 9186 has been "blamed for slow progress in the standardization of public information symbols." Increased reliability, simplified procedures and an increased likelihood that a test will yield an acceptable variant are all important considerations in a testing procedure, according to Brugger. Inclusion of this phase in the standards will further improve the likelihood that acceptable variants will result.

It is suggested that three variants be selected for testing to adjust for inconsistencies across cultures. In 1974, Easterby and Zwaga conducted a study of 3 ISO testing procedures. Originally, only the best symbol of this set was studied in detail by testing comprehension and applying a matching procedure. But in cross cultural evaluation of testing procedures, Easterby and Zwaga (1976) found inconsistencies across cultures, so the decision was made to select three variants.

A number of preliminary methods have been tested and recommended by ISO. The ANSI standards can benefit from this research. These methods are described more completely in ISO documents. (Brugger, 1994)

ISO has tested and recommends the use of the Preference Ranking Test.

Zwaga's (1989) results suggest that the preference ranking test could be replaced with estimation scores of comprehensibility, in some cases. In the estimation procedure, one referent and all its variant symbols are printed on one page in a circle. In the center of the circle are printed the name of the referent, its function and excluded functions. Next to each symbol, participants write an estimation of the percentage of the country's population that would understand its meaning. This method is highly predictive of comprehension scores. Zwaga suggests that the estimation test could eliminate the need for comprehensive testing of symbols that fall outside a 20% margin of error (i.e., those that fall on either side of the 67% ISO standard, below 47% and above 87%) If this method were adopted by ANSI, the higher ANSI standard of 85% leaves only a 15% error of margin above 85%, thus usually only symbols with a score of 65% or less could be eliminated, and all of the higher scores, or potential candidates, would have to be retested with final comprehension testing, unless the 85% rule were lowered or made adjustable. The margin of error is dependent upon the standard deviations in the scores, which will vary depending on participants and testing situations.

6. The complexity of visual information suggests an emphasis on qualitative information gathering

The complex nature of visual data demands a qualitative information gathering approach to evaluation. Comprehension testing provides this approach.

Repeatedly in symbols testing literature, the emphasis in evaluating testing methods is on the amount of information yielded. Tests are routinely rejected for yielding too little information for the work. For instance, Brugger states in his evaluation of testing methods, "The data of ... the matching test, *(mentioned with reservations in the 1991 ANSI guidelines* ... did not yield sufficient information to justify the high demands on the field work resources." (Brugger, 1994). The matching test, however, does provide information about confusions with other symbols.

Despite the appeal of quantitative methods, scientists in most fields that manipulate visual data have recognized the unavoidability of qualitative and subjective measures of evaluation for valid results. The understanding and manipulation of visual data requires subjective evaluation and qualitative methods to adequately explain and embrace all the factors. Visual data also needs quantitative and empirical methods to provide structure, perspective and a method of measurement. Both are needed for valid results.

Even computer system analysts in the area of visual languages have found it difficult to find ways to quantify description of pictures. Editor E. J. Golin of the Journal of Visual Languages and Computing writes in an editorial on Theory of Visual Languages,

> "Visual programming languages have largely been approached using ad hoc techniques... Visual languages are often specified intuitively,... by giving examples... or informal descriptions of the structure... The difficulties in specifying visual languages arise from the multi-dimensional nature of pictures. Pictures contain complex and diverse relationships between components, which require powerful mechanisms to describe."

7. A lack of theory on visual context suggests that experimental research be continued to derive some theory in this area.

The following insights were discovered too late to explore in this current work, but the suggestions are too valuable to be left out. Perhaps they will provide clues for future research, so they are included here.

Dr. George Miller, a well known, respected and seminal researcher in the area of memory and language spoke on the topic of contextualization at the 17th Annual Conference of the Cognitive Science Society in Pittsburgh, PA on July 24th, 1995.

Miller draws the tentative conclusion that there is some mechanism, not fully understood, whereby human beings can use context of a sentence or conversation to correctly understand the meaning of words most of the time. Thus far, no extractable rules are able to be derived and translated into a computer program to simulate that understanding. This paper again uses cognitive science and computer research as a benchmark of the strength and applicability of theory. If a theory holds water, a computer algorithm is often derivable from it. (For example, John Anderson's well-defined ACT-R theory.) When computers fail to mimic human understanding, it suggests that the theory is too complex to simplify or is not sufficiently understood.

Miller wrote in a later e-mail correspondence that this "talk was my first attempt to 'go public' with [this viewpoint]. I first encountered the problem when I tried to get computers to identify the context-appropriate sense of polysemous words, and I gradually became convinced that this was not language-specific, but represented a general cognitive ability that gives meaning to all experience." Miller further suggested that "the contextualization of visual icons is the same as for text". He wrote "your experience with visual icons is exactly parallel to mine with words". He wrote, "An icon is minimal, non-informative context (I am reluctant to say that an experimental context is no context at all) can have any of several meanings, just as a word can be polysemous. Different contexts lead people to select different meanings, or, to put it more carefully, context enables you to narrow the range of alternative possibilities."

Miller also pointed out that the lack of theory in this field of work [i.e. understanding how human's contextualize visually] "is probably representative of the difficulty of making progress--the data are interesting, but there seems to be no body of theoretical and experimental work to attach them to, no hypothesis outstanding that they can be used to test."

Miller made some suggestions which may be fruitful in further research on visual contextualization. "In the linguistic case, there are some hypotheses that we can test. E.g., is local context more important than topical context? It makes me wonder, what would correspond to this local/topical distinction in the realm of icons? As I think of the role photographs could play, they seem like topical contexts. Topical context, of course, is what children use to learn their language; local context becomes useful only later when word order becomes important and polysemous words [e.g. words with more than one meaning] pose a potential problem. You say your tests were short, so I assume

that your Ss had no chance to learn anything and were using either minimal context or topical context. Maybe some of the discussion about early vocabulary growth would provide some theoretical ideas for icon recognition?" (Miller, 1995)

In a second e-mail correspondence, Miller pointed out "Incidentally, someone else told me that work on scene recognition makes the distinction between topical and local visual context. In interpreting X-rays, for example, the topical context would be knowing what organ you were looking at; the local context has to do with tiny details of blood vessels and spots and cracks, etc. He said they had not expected the local context to be so important but were driven to it by their data." (Miller, 1995)

Researchers can look to the body of literature on verbal contextualization for some analogical parallels. Given the necessity of context in understanding language, are symbol experimenters asking to much to demand that a symbol be clear and unambiguous without providing appropriate contextual cues?

The language used in the Miller paper uses vocabulary specific to this narrow area of cognitive science, and prevents useful direct quotation, however, the meaning derived by this researcher is this: Without the use of any strategy, current computer programs written to understand the meaning of a noun, verb, adjective or adverb are only correct 45% of the time. When algorithmic heuristics are applied which use guessing, the most frequent meaning, and co-occurrence of other words are used, then the computer is correct 69% of the time. When only polysemous words (words with multiple meanings) are used, the computer is only correct 58% of the time. (Miller, Chodorow, Landes, Leacock and Thomas, 1994)

Since the ANSI criteria of 85% for comprehension of symbols is far higher than the best scores by computers in comprehending words, we can perhaps understand the existing difficulty of symbol researchers in reaching that criterion with many symbols. The specific response to this research is unclear at this time, but certainly, it may also support this paper's contention that context is critical to human comprehension and the use of it should be encouraged, and that the 85% criteria is perhaps unrealistic.

8. The 85% standard for comprehension is arbitrary

The 85% rule is arbitrary. World standards range from 65% to 85%, and the ISO standard is 67%. Different warnings are more important, a deadly hazard should perhaps have a higher criterion level.

Different kinds of symbols demand different criteria for assessment. Highway signs might need to be legible and understandable at a distance and when seen for a brief time and legible under adverse conditions such as glare, low light and poor weather. Fire Exit signs must be legible under adverse conditions such as smoke, low light, and should command attention and be easily detected.

Because of the complexity or abstract nature of some kinds of hazards, it may be impossible to develop an obviously comprehensible symbol. In these cases, (biohazard, laser hazard, etc.) other criteria such as learnability or discriminatability might be more important. In these cases, it is important that standard organizations establish some agreement on a proposed symbol so that consistent usage and public information campaigns can be realized world wide.

Dewar outlines a number of testing methods, "Psychological and psychophysical methods such as reaction time, glance legibility, signal detection, legibility distance, comprehension and preference ratings have been successfully employed to gauge the effectiveness of both existing and new symbols."

He states that "the relative importance of these various criteria has never been established". The criteria are "not all of equal importance and can be in conflict." In 1988,

Dewar contacted experts in four countries and they were all in agreement that comprehension was most important, followed by conspicuity, reaction time, legibility distance and learnability, in that order.

One of only a few studies to use a number of measures, Roberts et al. (1977) used five measures: 1) the time it took to understand the sign's meaning, 2) certainty or confidence in the participant's understanding (eliminates guessing), 3) comprehension, 4) preference ranking and 5) the minimum time needed for participants to accurately identify all the elements of the symbol. But the problem of assigning a percentage of importance to each of those measures was not addressed, as they were all rated equally. A lack of correlation indicated that most variables were measuring different factors in symbol effectiveness.

Dewar discusses the importance of keeping in mind the information system into which an icon will be placed so that when new symbols are introduced they are not confused with an older symbol. Dewar points out that the need to design highly discriminatable symbols goes counter to the notion of creating similar images to represent similar functions to fit with the Gestalt law of similarity, whereby human perceptual information processing tends naturally to group together visually similar elements.

In addition, different hazards require higher or lower standards for acceptance. Safety hazards causing loss of life, limb or expensive property might reasonably require a higher criterion than 85%. A safety hazard which is less dangerous could demand a lower criterion.

V. RESULTS & DISCUSSION

Context Effects

Symbols that had external environment detail in the symbol itself, showed no context effects. The simpler symbols without environmental clues showed strong context effects of an increase of 15 to 18 percentage points in the score. The artificially high scores in the low plausibility condition washed out differences between context and no context conditions, so that effects of only 7% were obtained.

Context eliminated wrong responses where an icon was clearly out-of-context. A possible explanation is that in the flammable hazard symbol, it was clear that the environments would not permit fires. For the eye protection symbol, context made it clear that eyeglasses were not being sold, and it was not an optometrist's office. In the hand protection symbol, no traffic cops, crossroads or deaf persons were in evidence.

Machinery in an actual environment may provide clues to discriminate between the nature of danger such as pressure release, flammable hazard, and explosives. The engineering student population may also have been knowledgeable concerning these hazards.

The Effect of Distractor Plausibility

The low plausibility distractors yielded invalidly high scores. Scores were directly related to the low plausibility of distractors. The Exit, Keep Door Open, and Sever Hazard symbols all obtained high scores in the original Mining study and their distractors obtained low plausibility scores.

The plausibility scores for the Mining study distractors were 2.5 on a scale of 7, below the average of 3.0 for all of the distractors tested. The median overall score for the Mining study distractors was 2.5 or halfway between “very poor match” and “poor match”, while the median score for plausible distractors obtained in the 1995 Wolff

study was 4.9, (excluding the Mining study distractors) or "good match". The average of all distractors (including the Mining study distractors) used in the plausible condition was 4.0, and the average of the distractors used in the low plausibility condition was 2.0. The average of the discarded distractors was 3.0. The average of the correct answers was 4.9.

The average score of all 33 symbols tested in the condition with low plausibility distractors was 88% compared to 59% for symbols with plausible distractors, a difference of over 30 points. The open ended score averaged 55%, just 4 points under the multiple choice method with plausible distractors.

The failure to identify just one plausible distractor resulted in passing grades for poor pictorials in the multiple choice condition for, No Entrance, Crush hazard, Sever Hazard, and Pressure Release.

When the multiple choice method worked.

The multiple choice method equalled the comprehension method only with good pictorials or pictorials that had small numbers of confusions.

When the multiple choice method did not work

Symbols with a large number of confusions i.e.. Keep Door Closed, Exit, No Entrance received invalidly high scores by multiple choice testing, as did symbols where there were no plausible distractors possible at all, i.e. Eyewash Located Here.

One symbol, Keep Door Open, had 4 highly plausible distractors. Obviously, since only 3 distractors are allowed, one of the distractors had to be discarded. One of the distractors that had to be left out, "Caution, Swinging Door," occurred with a higher frequency in the open-ended test, thus suggesting that the score of the multiple choice test was invalid. This example shows that even when difficult and time-consuming effort is made to obtain plausible distractors, the simple limit of 3 distractors can prevent a valid score.

Only the comprehension method correctly identified these problems

The Poison Hazard (#22) has a poorly written referent which does not match the symbol. This can be seen in the low plausibility score given the correct referent (3.4). The symbol is clearly picturing an airborne poison hazard or fumes, not a liquid or solid poison. This became clear in judging because the judges were only given the written results and the written referent to compare, not the symbol. The use of this incorrect referent with the symbol could result in its being used for a non-airborne poison hazard, with inadequate protection.

A rare perceptual confusion is seen in the Eyewash Located Here symbol (#12). Participants could not identify the face and fountain pictured in the symbol. The written distractor on a multiple choice method invalidly gives away the meaning of the symbol in the test situation. This would not occur in a real life situation.

Some participants recognized the eyewash apparatus in the context condition of the test. This context effect can be seen in the open-ended condition, but not in the multiple choice condition, because the distractors already provided that answer.

According to Cairney and Sless, two kinds of confusion may arise. At the perceptual level, there can be wrong recognition of the symbol content, and cognitive confusions occur when the image is recognized, but the wrong meaning attributed to it.

Perceptual confusions were rare but it is unlikely they will be identified in any test with written answers for selecting.

The Importance of context in testing simpler symbols.

One typical problem in comprehension is with a literal transformation of symbols (Dewar, 1994) The referent is more general, but the symbol is too restrictive in its interpretation. Usually, a too specific symbol has contextual information designed into the symbol. ANSI recommends simple symbols with little distracting detail. These tend to be understood more generally, but also are more susceptible to out-of-context interpretations. Thus, it is critical that when these symbols are tested, some photographic context is provided to eliminate these answers that bring comprehension scores down.

Issues in choosing photographic context

Although providing photographic context in a testing situation is important, choosing one or more photographs presents a number of issues.

A photograph is not benign, as any photographer knows, a photograph, by its cropping, lighting, angle and focal length can imply danger or emphasize certain objects in the photo. Also, the choice of image makes a difference.

The photographs in this study were not ideal, but in the process of choosing photographs and observing participant responses, some initial guidelines were identified. A photograph should show an environment, rather than a person. If a person is shown, they should not be shown specifically engaging or not engaging in the prohibited or suggested behavior. This could confuse or bias the test participant. An example is one photograph of a person using a corrosive without gloves used to test "wear safety gloves". A better photograph might have been a photo of the corrosive in the bottle and the work site. Another example is the photograph of people digging used for the "do not dig" symbol. This may have implied to participants that digging was allowed, and therefore the symbol meant something about how or where to dig. This was the only symbol tested where the scores for the context condition were actually lower than the no-context condition. The symbol already contains a great deal of contextual detail, so the expected result would be no difference in the context condition. This was obviously a poor choice of photograph.

It is possible that the validity of a rating for one symbol could only be determined for those contextual situations where it had been tested. In other words, a symbol would be approved with reservations. If a particular industry wished to use a symbol, they would be able to identify in which contexts it had been tested successfully.

Should All Symbols Contain Contextual Detail?

Symbols with contextual detail may score better than simpler symbols in a testing situation because of a lower ambiguity. In a real world environment, where the environment provides the context, a simpler symbol may perform better. In any case, it is useful to be aware of the pros and cons of each. It may help to explain test results and indicate important information to consider when testing a symbol.

Although the right kind of contextual detail may make a symbol score better in the open-ended test, but one should not therefore assume that the more detail, the better. Contextual detail also makes a symbol more confusing. It may be harder to recognize at a distance, and difficult to recognize in a small size. Also, a more complex symbol may take more iterations to find a good combination of details.

Future Issues and Problems

No really abstract symbols were included in the symbol set, with the exception of the explosion symbol. They should be included in the future studies. It is hypothesized that the testing of these symbols would be particularly problematic with the multiple choice method because of the range of answers that might be obtained.

Further studies need to be done that exploit multimedia technology . Authoring software for testing symbols could be developed and provided by an independent contractor or ANSI.

A World Wide Web site could be established for facilitating the collection, distribution and dissemination of symbols and symbol standards world wide. Artificial intelligence software for the search and recognition of symbols in a world wide symbol database may be one of the only ways to facilitate the thorough gathering of all existing symbols matching a particular referent.

The current study attempted to cover important issues with a broad brush. Additional studies need to look at the difference between verbal, descriptive context and photographic, pictorial context in testing situations.

Another study could look at the different kinds of problems and issues in choosing photographic context images. For instance, use of persons in the photograph, depicting the prohibited or recommended action. Cropping, angle, and colors in photos may all affect the clues provided and participant response. It is unclear whether providing more than one photograph gives an unrealistic simulation of a real world environment. It was clear that a single picture may have provided unintended messages. It was hypothesized that multiple photographs minimized the effect of any particular detail in the photographs and provided the more general effect that a time continuous, three dimensional world without limits and cropping would provide. Another test could manipulate this variable to test this assumption.

It is clear that learning from news reports and safety and informational ad campaigns affects the interpretations of participants. Anti-drug, bottle tampering, anti-drinking and driving and child protection slogans all appeared in written answers. Such slogans were often substituted when a symbol was confusing. In addition, learned responses and interaction with symbols such as the men's room door symbol affected the understanding of a similar symbol such as the “No entrance” symbol. It is possible that the overuse of graphic guidelines may result in overly similar symbols which lose their ability to communicate because participants cannot discriminate between similar variants. Symbols may take on an unintended meaning by their use in a particularly common symbol.

VI. REFERENCES

Boersema, T., and Zwaga, H. J. G. (1989). Selecting comprehensible warning symbols for swimming pools. In *Proceedings of the Human Factors Society 33rd Annual Meeting.* (pp. 994-997.) Santa Monica, CA: Human Factors Society.

Brugger, C. (1994). Public information symbols: A comparison of ISO testing procedures, In *Proceedings of Public Graphics,* 26-30 September, (pp. 26.1-26.10.). Lunteren, The Netherlands.

Cahill, M.C. (1975). Interpretability of Graphic Symbols as a Function of Context and Experience Factors, *Journal of Applied Psychology*, *60*(3), 376-380.

Cairney, P., and Sless, D. (1980). Understanding symbolic signs: design guidelines based on user responses. In *Proceedings of the 17th Conference of the Ergonomics Society of Australia and New Zealand*. (pp. 51-58).

Collins, B., (1983). *Use of Hazard Symbols/Pictorials in the Minerals Industry*, National Bureau of Standards, Bureau of Mines, Washington, D.C.

Dewar, R., Arthur, P. (1994). Warning of Water Safety Hazards with Cartoon Images, *Proceedings of Public Graphics*, 26-30 September, (pp. 26.1-26.10). Lunteren, The Netherlands.

Dewar, R. (1994). Design and Evaluation of graphic symbols, In *Proceedings of Public Graphics*, 26-30 September, (pp. 24.1-24.18). Lunteren, The Netherlands.

Dunlap, G. L., Granda. R. E., Kustas. M. S. (1986*). Observer Perceptions of Implied Hazard: Safety Signal Words and Color Words*, IBM Technical Report, TR 00.3428.

Goldhaber, G. M., de Turck, M. A. (1988). Effectiveness of Warning Signs: Gender and Familiarity Effects, *Journal of Products Liability, 11*, pp. 271-284.

Golin, E. J. (1991). "Theory of Visual Languages," *Journal of Visual Languages and Computing,* 2(4), Harcourt, Brace and Javanovich, pp. 309-310.

Grisim, J. T. (1993). *Product Warnings - Practical Implications of ANSI Z535 Standards on Product Labeling,* Presentation to National Safety Council Congress & Exposition.

Keeler, M.A., and Denning, S.M. (1991). The challenge of interface design for communication theory: From *Interaction metaphor to contexts of discovery. Interacting with Computers*, *3*, 283-301.

Kraemer, H.C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage Publications.

Lewis, C., & Norman, D.A. (1986). Designing for error. In D. A. Norman & S. W. Draper (Eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction* (pp. 411-432). Hillsdale, NJ: Lawrence Erlbaum.

Lewis, J. R. (1994). Sample Sizes for Usability Studies: Additional Considerations. *Human Factors*, *36*(2), 368-378

Nielson, J. (1993). *Usability Engineering*, London: Academic Press.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of ACM INTERCHI'93 Conference 24-29 April,* (pp. 206-213). Amsterdam, the Netherlands.

Silver, N. C., Wogalter, M. S., Magurno, A. B., & Glover, B. L. (1995). Comprehension and Perceived Quality of Warning Symbols. In *Proceedings of the Human Factors Society 39th Annual Meeting.* Santa Monica, CA: Human Factors Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation. How many subjects is enough? *Human Factors*. *34*(4), 457-468.

Virzi, R. A. (1990). Streamlining the design process: running fewer subjects. In *Proceedings of the Human Factors Society 34th Annual Meeting* (Vol. 1, pp. 291 - 294). Santa Monica, CA: Human Factors Society.

Wolff, J. S., & Wogalter, M. S. (1993). Test and development of pharmaceutical pictorials, In *Proceedings of the Interface '93,* (pp. 187-192). Santa Monica, CA: Human Factors Society.

Wogalter, M. S., Wolff, J. S., Magurno, A. M., & Kohake, J. R. (1994). Iterative Test and Development of Pharmaceutical Pictorials, *Ergonomics and Design*, IEA '94, 4, 360-362.

Zwaga, H. J. (1989). Comprehensibility estimates of public information symbols their validity and use, *Proceedings of the Human Factors Society 33rd Annual Meeting*.

Note: Not all of the Appendix materials are available in postscript files. Thus, the appendix is only partially complete. However, they are available in paper form from the GVU library and from the Georgia Tech Library.

APPENDIX A

ANSI (American National Standards) Z535.3-1991

Criteria for Safety Symbols, Annex A (normative),
Suggested procedure for evaluating candidate symbols

(Available from NEMA, National Electrical Manufacturer's Association in Washington, D.C.
This appendix is included in the paper version of this GVU Technical Report)

APPENDIX B

Submission to ANSI committee on Z535

The following 11 pages were submitted to the ANSI Committee revising the Z535 standard.

These pages were forwarded to the 50 member committee for consideration. Several committee members requested the full copy of the study.

(This is essentially an abreviated form of the text of the paper.
This appendix is included in the paper version of this GVU Technical Report)

APPENDIX C

Final Results for 33 Safety Symbols Tested

(Appendix C is a separate postscript file, available online)

APPENDIX D

Test Materials for Part 1

1. A Sample Page from the Plausibility Rating Test Booklet

2. Open-ended Test Booklet from the Plausibility Rating Stage

(This appendix is included in the paper version of this GVU Technical Report)

APPENDIX D (Continued)

Test Materials for Part 3

1. Instruction Sheets

2. Biographical Data Sheet

3. Georgia Tech Consent Form

4. Sample Page from the Multiple Choice Answer Sheet
(High and Low Plausibility Distractors)

5. Answer Sheet for Both Open-ended Test Conditions

6. Open Ended /No Context Test Booklet

(This appendix is included in the paper version of this GVU Technical Report)

APPENDIX D (Continued)

Test Materials for Part 3

7. Color Xerographic Context Booklet

(This appendix is included in the paper version of this GVU Technical Report)

APPENDIX D (Continued)
Sample Judge's Scoring Sheets

All responses in the Context Condition are numbered in the 400's

All responses in the No Context Condition are numbered in the 300's

Both 300 and 400 numbers were coallated in numerical order and given to the judges together to avoid effects of learning and fatigue.

Each page of the participant's responses were coded individually for tracking purposes.

(This appendix is included in the paper version of this GVU Technical Report)

APPENDIX E

Previous Published Papers by J. S. Wolff

1. Iterative Test and Development of Pharmaceutical Pictorials
2. Test and Development of Pharmaceutical Pictorials

(This appendix is included in the paper version of this GVU Technical Report)